

# Exploring crowdsourcing information to predict traffic-related impacts

Pavlos Tafidis, João Texeira, Behnam Bahmankhah, Eloísa Macedo, Margarida C. Coelho, Jorge Bandeira

Department of Mechanical Engineering, Centre for Mechanical Technology and Automation

University of Aveiro

Aveiro, Portugal

pavlostafidis@ua.pt, jpteixeira@live.ua.pt, behnam.bahmankhah@ua.pt, macedo@ua.pt, margarida.coelho@ua.pt, jorgebandeira@ua.pt

**Abstract**— Due to the increased public awareness on global climate change and other environmental problems, advanced strategies and tools are being developed and used to reduce the environmental impact of transport. The main objective of this paper is to explore the potential of using crowdsourcing information as an alternative or complementary source data to predict traffic-related impacts. Three main road connections to two important commercial areas in the city of Aveiro in Portugal, are examined. Driving patterns over different periods were collected using a probe vehicle equipped with a GNSS data logger and volumes of traffic were counted during different days of the week. The emissions estimation was based on the concept of Vehicle Specific Power (VSP), which has the capability to predict emissions during a trip often-according recorded second-by-second vehicle dynamics. Various tests were conducted in order to explore the potential correlations between these data sets and the information of a certain place's busy times that are provided by Google Maps. The findings of the study prove the potential of crowdsourcing information and shows that ICT technologies can be used to estimate environmental and traffic-related impacts.

**Keywords**—crowdsourcing; traffic impacts; air pollutants; emissions

## I. INTRODUCTION

The increasing road transport volumes in urban areas are a primary source of air pollution and greenhouse gas emissions [1]. The conditions are usually worst in areas that generate and attract many trips, such as city centres, shopping areas etc. In some cases, the low level of public transport services or even more the absence of alternative transport modes encourage the use of private vehicles aggravating the situation even more.

Advance traffic management systems are essential tools to reduce air emissions and promote sustainable mobility. They

---

The authors acknowledge the support of Project CISMOB (PGI01611), funded by the Interreg Europe Programme and the Strategic Project UID-EMS-00481-2013. M. C. Coelho and B. Bahmankhah also thank @CRUiSE project (PTDC/EMS-TRA/0383/2014), funded within the Project 9471 – Reforçar a Investigação, o Desenvolvimento Tecnológico e a Inovação (Project 9471 – RIDTI) and supported by the European Community Fund FEDER.

integrate multiple technologies to improve traffic flow and reduce congestion. High quality real-time traffic data is necessary to maintain the effectiveness and reliability of traffic management services. In transport systems, traditional data collection systems is usually considered costly and lengthy and limited to specific areas of interest. Many initiatives tried to estimate traffic and environmental impacts using new sources of information. For a case study in Chicago, a methodology was developed based on Bing Maps information to estimate traffic [2]. Furthermore, several studies have been conducted also regarding the potential of using data from various social media for transport issues [3,4,5].

In the last few years, plenty of new applications are using crowdsourcing to give users real-time information through graphs or photos about the traffic of specific locations. For instance, Google launched in July 2015 a new feature in Google Maps that allows users to know a certain place's busy times by crowdsourcing users' visits. The new update aims at helping people in their decision-making process about when it is the best time to visit a business, a restaurant, a shop etc. The data usually includes: i) popular times graph per day and hour showing how busy is a specific location based on average popularity over the last several weeks, ii) live activity data updating by real-time information and easily comparable with the average values and, iii) visit duration, showing the average spending time of people at the place. This sort of information could be especially useful as an alternative source of predicting traffic related impacts in cities or areas where floating car data (FCD) penetration rates are insufficient or traffic-monitoring stations are limited.

The aim of this paper is to explore an alternative source of data by examining if there is any correlation between the information provided by applications like Google Maps regarding the popular times of specific areas and the amount of air emissions produced during that period.

## II. METHODOLOGY

### A. Case Studies

In this paper, six road links are examined near two important commercial areas in the city of Aveiro, Portugal. Links L1 and L2 consist of the main entrance and exit to The

Aveiro Shopping Center and links L3 and L4 connect the city of Aveiro with the industrial zone (Fig. 1). The area is characterized by industrial land use, the studied links are between roundabouts and there are various unsignalized intersections with minor roads. Regarding the second study area, link L5 is the main street that lead to Centro Comercial Glicínias Plaza, while link L6 leads out of it (Fig. 2). The land use of the area is mainly residential and the studied links are between two-lane roundabouts. One crosswalk interrupts them, while in link L6 there is the only entrance and exit to a gas station. To enhance the applicability of the proposed methodology, we chose links that clearly serve as traffic distributors to the commercial zones but also, due to their proximity to important arterials, the traffic performance can be affected with traffic from different Origin–Destination pairs.



Fig. 1. Study area of Aveiro Shopping Center

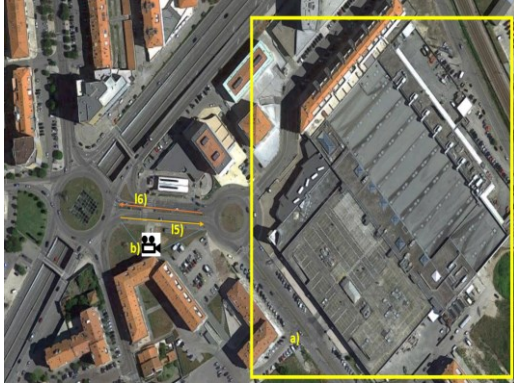


Fig. 2. Study area of Centro Comercial Glicínias Plaza

## B. Data Collection

For the purpose of the paper three different data sets were collected:

- Traffic volumes in 15 minutes intervals with the use of cameras.
- Traffic dynamics (travel time, speed, and acceleration) with the use of a light-duty vehicle equipped with a GNSS data logger to collect second by second trajectory data required for microscopic analysis.

- Crowdsourcing information in real time (Popular Times) from Google Maps regarding the activity of the commercial areas (Fig. 3). Google Maps present Popular Times with the use of a bar chart without providing values, For the purpose of this paper, we assume that the minimum value of the bar is zero and the maximum is one and we divided it in ten equal parts giving them the respective values.



Fig. 3. Example of popular times feature of Google Maps

The data collection period was chosen in order to obtain a diversified range of demand in both case studies, including normal weekday and special weekend conditions and lasted 6 hours each day. Regarding the vehicle dynamic monitoring, the number of 10 runs were performed per hour for each link by different drivers to increase the heterogeneity of driving behavior [6]. The probe vehicle was moving according to the driver's perception of the average speed of the traffic stream [7].

A MATLAB routine developed by one of the authors to organize data and calculate dynamic data obtained during field tests and write them in Excel was used. Excel was used to perform data exploratory analysis, mainly focused in establishing associations between variables.

## C. Emission Estimation

The Emissions estimation was based on the concept of Vehicle Specific Power (VSP). The VSP model reflects the comparison of driving behavioral effects in fuel consumption and vehicle emissions for different air pollutants:

- Carbon Dioxide (CO<sub>2</sub>).
- Carbon Monoxide (CO).
- Nitrogen Oxides (NO<sub>x</sub>).
- Hydrocarbon (HC).

The VSP represents the power required to the engine based on the road gradient, aerodynamics, kinetic energy and friction to the movement and it is a model that has proven to be very effective in estimating emissions from petrol and diesel cars [8]. The VSP mode can be expressed as

$$VSP = v[1.1a + 9.81(\tan(\sin(\text{grade}))) + 0.123] + 0.000302v^3 \quad (1)$$

where:  $v$  = vehicle speed (m/s),  $a$  = vehicle acceleration/deceleration rate ( $m/s^2$ ), grade = vehicle vertical rise divided by the horizontal run (%). Each VSP bin refers to one of 14 modes. Each VSP mode is defined by a range of VSP values which are associated to an emission rate. Each calculation of VSP results in a unique classification to a VSP mode [8,9]. The following distribution fleet composition was considered based on the Portuguese vehicle classification for the case study: 38% of light duty gasoline vehicles and 62% of light duty diesel vehicles [10]. It should be noted that due to the flat terrain, the grade of road was considered negligible.

### III. RESULTS

TABLE I displays the main characteristics of each link after the analysis of the results. Fig. 4 presents the two approaches that were followed to find the relationships between the traffic volumes and the Popular Times of the commercial zones. The standard interval of 15 minutes was used in this study as it is the lowest interval which flow rates are statistical stable [10]. In the first approach, traffic volumes of each interval were compared to popular times' value in the end of the respective time period, while in the second the comparison was made with the value in the end of the next 15 minutes period. Although the tests using the second approach did not provided significant improvement on the results obtained for the first, we only report the first approach. The main objective of this study is to explore the potential of crowdsourcing information to contribute as an alternative source of data for real-time traffic estimation and traffic-related impacts. The findings of the study prove the potential of crowdsourcing information.

TABLE I. CHARACTERISTICS OF LINKS

Links	Length (m)	$Q_{max}$ (volume per hour)	Average Q (volume per hour)
L1	215	816	544
L2	210	924	608
L3	718	210	210
L4	722	210	210
L5	88	700	700
L6	90	700	700

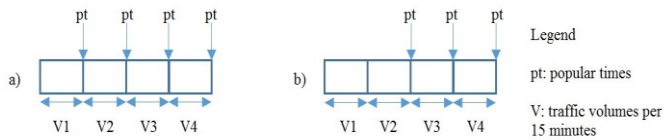


Fig. 4. a) First (a) and second (b) approach for corellation analysis

#### A. Centro Comercial Glicinias Plaza

##### 1) Relationships between various variables and Popular Times

To study the relationships between traffic volume, travel time, emissions and Popular Time, we try to use simple models, such as linear and quadratic fittings, in order to ease the interpretation of results. We are interested in whether Popular Time, as a predictor variable, can explain the variability of other variables. From Table II, the first observation is that the results for the first day are not so encouraging, while the second day presented better significance values (p-value). This means that the chosen models are adequate and variables are correlated with Popular Time, i.e., changes in Popular Time (predictor) are associated with changes in the other variables (response). Considering both days results, we can see that Traffic Volumes and CO<sub>2</sub> emissions are the variables that Popular Time can explain better, since they present highest coefficient of determination values. Better results are reported in second day for Link 6, where we can see that Traffic Volumes is highly linearly correlated with Popular Time, concretely, 90% of the variability of Volumes can be explained by Popular Time. We can also find that more than 61% of the variation in CO<sub>2</sub> emissions, and more than 51% in Travel time and NO<sub>x</sub> emissions, can be explained by Popular Time. These results support the idea that ICT applications can be used to predict traffic-related impacts.

TABLE II. MODEL PARAMETERS BETWEEN RESPONSE VARIABLES AND POPULAR TIME

Links	Variables	Coefficients of the model			R <sup>2a</sup>	P-value
		<i>a</i>	<i>b</i>	<i>c</i>		
1 <sup>st</sup> day						
L5	Volume	2622.9	-2059	908.75	0.21	3.78E-04
	Travel Time	4504.6	6502.3	linear fitting	0.04	1.21E-01
	System CO <sub>2</sub>	6744.9	11556	linear fitting	0.05	6.88E-02
	System NO <sub>x</sub>	15.016	25.805	linear fitting	0.05	7.42E-02
L6	Volume	342.32	417.73	linear fitting	0.44	2.17E-10
	Travel Time	162709	-121358	36181	0.12	1.31E-02
	System CO <sub>2</sub>	139586	-99243	39538	0.13	7.63E-03
	System NO <sub>x</sub>	374.81	-270.05	108.48	0.12	1.27E-02
2 <sup>nd</sup> day						
L5	Volume	4933.6	-3372.2	881.91	0.44	3.14E-10
	Travel Time	77890	-50449	12499	0.41	2.70E-09
	System CO <sub>2</sub>	76797	-42000	12741	0.41	2.50E-09
	System NO <sub>x</sub>	120.92	-51.293	20.812	0.34	1.71E-07
L6	Volume	1372.7	-177.29	linear fitting	0.91	7.78E-42

Links	Variables	Coefficients of the model			R <sup>2a</sup>	p-value
		a	b	c		
	Travel Time	37315	-6703.2	linear fitting	0.52	5.63E-14
	System CO <sub>2</sub>	55962	-10426	linear fitting	0.61	1.34E-17
	System NO <sub>x</sub>	158.1	-29.962	linear fitting	0.52	1.87E-13

a. Coefficient of Determination

## 2) Correlations between Traffic Volumes and Popular Times

Fig. 5 presents the correlation between traffic volumes per hour (vph) near Centro Comercial Glicínias Plaza and popular times for the two days.

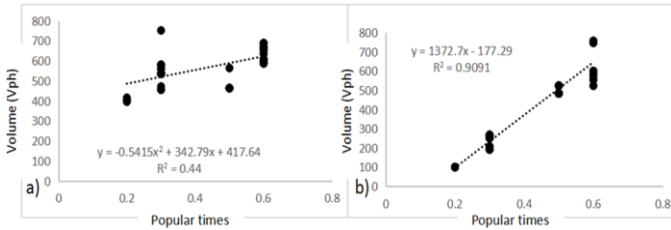


Fig. 5. a) Quadratic correlation for link L6 in first day b) Linear correlation for link L6 in second day

For the sake of clarity, we are aware that no accurate assertion can be made in terms of clear correlations between such variables when values are registered at the same time. Evidently, one cannot say that in the moment we registered a vehicle, the people on it are already inside the shopping. Nevertheless, we believe that there is, however, some relationship between the traffic volume and the volume of people inside the shopping center within a delay of few minutes.

In this light, the first observation we can make from Fig. 5 is that during the weekend presented better results than the weekday. Concretely, in the second day, there can be verified a linear fitting, but we can improve the correlation coefficient with a third-order polynomial fitting. Regarding the first day, no clear correlations can be presented with the same fittings. From the results, we can assume that higher percentage of vehicles during the second day had as origin or destination the commercial area.

## 3) Correlations between Emissions and Popular Times

Fig. 6 shows the correlation between System CO<sub>2</sub> emissions (where it considers the Portuguese vehicle classification fleet and the traffic volume registered in each day) and System NO<sub>x</sub> emissions with popular hour for each day.

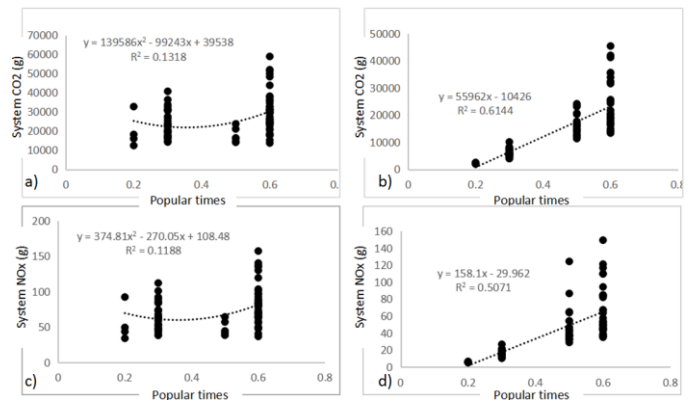


Fig. 6. a) c) Quadratic correlation for NO<sub>x</sub> and CO<sub>2</sub> for first day in Link L6; b) d) Linear correlation for NO<sub>x</sub> and CO<sub>2</sub> for second day on Link L6

There is no linear correlation on the first day, for the same reason that was mentioned above. In this case, the emission values are too high with respect to Popular Times values, which are between 0.2 and 0.3 where traffic isn't supposed to be so high. Considering the results on the second day, we can observe that there is higher correlation between emission values and Popular Times. This means that under similar circumstances Popular times can be a valid explanatory variable to predict road traffic emissions.

## B. Aveiro Shopping Center

### 1) Relationships between various variables and Popular Times

In Table III, the first observation is that we have better results, not only in terms of the coefficient of determination value, but also in terms of significance, when compared to Glicínias. In particular, we were able to estimate linear fittings for almost all cases. We can see that for links 1, 2 and 3, the estimation of the variability of NO<sub>x</sub> emissions by Popular Time is smaller than that obtained for the other response variables, nevertheless, we can assert that more than 38% of NO<sub>x</sub> emissions can be explained by Popular Time. Considering all the links, we can see that Traffic Volume and CO<sub>2</sub> emissions are the response variables that Popular Time can explain better. In particular, Popular Time can explain approximately 67% of Traffic Volumes and CO<sub>2</sub> emissions in Link 1. Concerning Link 2, 80% and almost 62% of the variability of Traffic Volumes and CO<sub>2</sub> emissions, respectively, can be explained by Popular Time, while in Link 3, such values are closer to 71% and 59%, respectively. For Link 4, we see that more than 50% of the variability of all response variables can be explained by Popular Time. Once again, the results show that we can estimate traffic-related impacts by using ICT applications.

TABLE III. COEFFICIENT OF DETERMINATION BETWEEN EMISSIONS AND TRAVEL TIME VARIABLES

Links	Variables	Coefficients of the model			R <sup>2</sup>	p-value
		a	b	c		
L1	Volume	942.31	85.737	linear fitting	0.67	8.74E-16



Links	Variables	Coefficients of the model			R <sup>2</sup>	p-value
		a	b	c		
	Travel Time	16284	-142.68	linear fitting	0.5	1.61E-12
	System CO <sub>2</sub>	27173	2950.1	linear fitting	0.67	6.41E-16
	System NO <sub>x</sub>	75.802	10.962	linear fitting	0.38	1.14E-07
L2	Volume	1097	74.875	linear fitting	0.80	8.47E-23
	Travel Time	18914	315.54	linear fitting	0.55	3.99E-12
	System CO <sub>2</sub>	33287	4141	linear fitting	0.62	4.07E-14
	System NO <sub>x</sub>	-185	280.97	-26.94	0.42	1.05E-07
L3	Volume	1005.6	-73.07	linear fitting	0.71	4.69E-17
	Travel Time	39230	-1.9003	4019.2	0.56	7.59E-11
	System CO <sub>2</sub>	137570	-15920	linear fitting	0.59	8.22E-13
	System NO <sub>x</sub>	341.72	-30.624	linear fitting	0.44	8.06E-09
L4	Volume	1051.4	-13.138	linear fitting	0.57	1.54E-12
	Travel Time	75585	-7482.7	linear fitting	0.52	7.57E-11
	System CO <sub>2</sub>	142103	-10772	linear fitting	0.60	1.98E-13
	System NO <sub>x</sub>	393.23	-23.937	linear fitting	0.55	7.82E-12

It can be pointed out that for Popular Times values between 0.2 and 0.6, it was drivers behavior that most influenced the level of emissions, while for values between 0.7 to 0.8, it was the travel time. It should also kept in mind that when there are registered low levels of Popular Times values, it could mean that more people is leaving the shopping area, thus, increasing the existing traffic and justifying the relation with drivers behavior. As well as in the case of higher Popular Times values, the weight of travel time is explained on the fact that more people are coming in than going out, meaning less vehicles exiting the commercial area.

## 2) Correlations between Traffic Volumes and Popular Times

Fig. 7 shows the correlations between traffic volumes and Popular values for the Aveiro Shopping Center. It is important to notice that the Popular Times value 0.5 there was a high variation of traffic volumes in all links. Link L3 and L4 offer one lane per direction, while link L1 and L2 offer two lanes per direction. In case of L1 and L2, that represent the main entrance to the commercial area it was expected a high correlation. In all links we can improve the correlation coefficient with a third-order polynomial fitting.

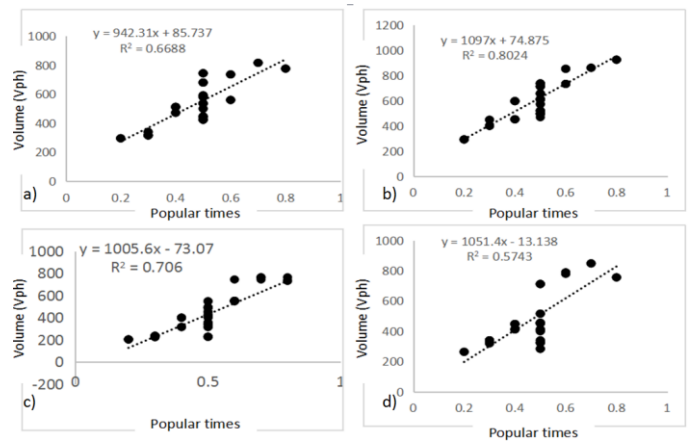


Fig. 7. a) Link L1; b) Link L2; c) Link L3; d) Link L4

## 3) Correlations between emissions, travel time and Popular Times

Fig. 8 displays low correlation between emissions and travel time with Popular Times compared with the case of Centro Comercial Glicínias Plaza in second day. These results can be explained by the drivers behaviour, which was shown to be inconsistent and aggressive in case of L1, thus producing more emissions, meaning that without traffic congestion, drivers were able to experience greater accelerations and speeds, such as greater speed reductions near the roundabouts. With some traffic, the drivers changed their behavior and begun to adopt a median driving style. We can also refer that the higher slope of the fitting, for low popular hour, is concerning the emissions. Popular Times valued 0.5, where the drivers changed their driving style, is a key point with major impact on the correlations. If we neglect this point, correlations for CO<sub>2</sub> go up to 0.79 and NO<sub>x</sub> to 0.67. In L1 driver's behavior was the most important factor (in both cases, with or without traffic).

Link L3 gave correlations lower than with L1. This is essentially because of Popular Times valued 0.8 that also appears in the peak hours at late afternoon, where many cars in L3 were mostly entering or exiting from Aveiro (once again it is important to refer that in this link driver's behaviour is strongly related to traffic volume). If Popular Times valued 0.8 is taken out of the correlations, then CO<sub>2</sub> correlation values will be 0.71, NO<sub>x</sub> will be 0.50.

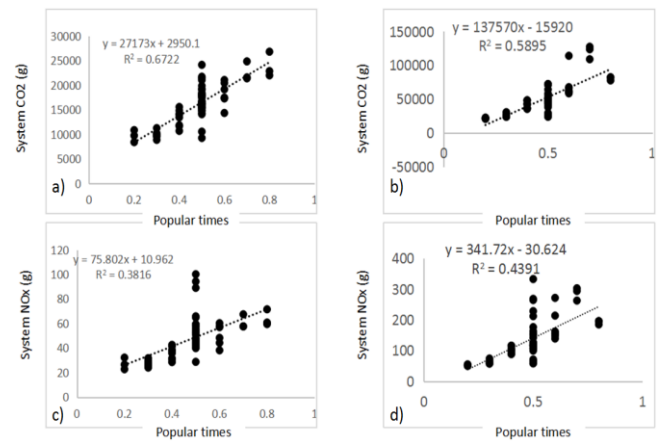


Fig. 8. a) P-value = 0.09; b) P-value = 0.24, correlations between emissions and popular hours; c) P-value = 0.14, correlation between travel time and emissions

#### IV. CONCLUSIONS

Certain levels given by google have less tests made than others. This implies that, in the gama of each of those levels, the conclusions or observations are dependent of the low population studied in those cases. This happens essentially in popular hour 0.2 or above 0.7, because the variations in this cases are very swift. In Aveiro Shopping Center the following was noticed: in every test, except L4 correlations between volume and popular hour, for linear and polynomial regressions were above 0.65. This happens because in this link the majority of road users use it to get out of the city and not go to commercial area. During the data analysis it was seen that good correlations between emissions and popular hours were seen when the driver adapted his driving style according to traffic, except for L1 because of the link characteristics, which makes sense giving the good relations between volume and google popular hours. Data was organized per each level given by google. It was seen that google levels also have some correlations with travel time. Another test was made but not presented, where it was considered that what we were measuring is not being related to what google is showing in the same time. In some cases results have proven better in the other they kept the same. With this in mind, it was proven the relation and potential to use crowdsourcing information to describe links nearby or emissions. The values are promising, however there is indeed a need for a better calibration between the time data is collected and the time where google supplies the information. The presented results support our goal, showing that ICT technologies/applications can be used to estimate environmental impacts.

#### REFERENCES

- [1] R. Colville, E. Hutchinson, J. Mindell and R. Warren, "The transport sector as a source of air pollution," in *Atmospheric Environment*, vol. 35(9), pp. 1537-1565, March 2001.
- [2] A.I.J. Tostes, F. de LP Duarte-Figueiredo, R. Assunção, J. Salles and A.A. Loureiro, "From data to knowledge: city-wide traffic flows analysis and prediction using bing maps," in the *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, pp. 12, August 2013.
- [3] M. Ni, Q. He and J. Gao, "Using social media to predict traffic flow under special event conditions," 2014 [Transportation Research Board 93rd Annual Meeting].
- [4] F. Yang, P.J. Jin, X. Wan, R. Li and B. Ran, "Dynamic origin-destination travel demand estimation using location based social networking data," 2014 [Transportation Research Board 93rd Annual Meeting].
- [5] E. Chaniotakis, C. Antoniou, J.M.S. Grau and L. Dimitriou, "Can Social Media data augment travel demand survey data?," [IEEE 19th International Conference on Intelligent Transportation Systems pp. 1642-1647, November 2016].
- [6] R. Downing, A. Skabardonis, V. Alexiadis, "Traffic Analysis Toolbox Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software," 2004.
- [7] S.M. Turner, W.L. Eisele and R.J. Benz, "Travel time data collection handbook," March 1998.
- [8] M.C. Coelho, H.C. Frey, N.M. Rouphai, H. Zhai and L. Pelkmans, "Assessing methods for comaring emissions from gasoline and diesel

light-duty vehicles based on microscale measurements," in *Transportation Research Part D: Transport and Environment*, vol.14(2), pp. 91-99, 2009.

- [9] H.C. Frey, A. Unal, J. Chen, S. Li and C. Xuan, "Methodology for developing modal emission rates for EPA's multi-scale motor vehicle & equipment emission system," Ann Arbor, Michigan: US Environmental Protection Agency, 2002.
- [10] Emisia SA., COPERT data, Retrieved April 9, 2017, from <http://emisiasa.com/products/copert-data>.
- [11] D. Florea, D. Covaciu, I. Preda and J. Timar, "Traffic counts methodologies for various type of road traffic applications".